北京大学中国经济研究中心
China Center for Economic Research
讨论稿系列
Working Paper Series

No.C2018013 2018-11-15

文本大数据分析在经济学和金融学中的应用：一个文献综述

沈艳 、陈赟、黄卓
北京大学国家发展研究院

摘要

本文对文本大数据分析在经济学和金融学中应用的文献进行综述。文本大数据具有数据来源多样化、数据量增长快和时频高的特征，能够为经济学和金融学研究提供新的信息和独特的分析视角，但是处理文本大数据存在许多挑战。我们梳理了文本大数据的信息提取步骤，总结了常用的词典法、机器学习方法和深度学习方法的实现原理和技术特点。在经济学研究领域中，文本大数据主要用于测度经济和政策不确定性、基于文本的行业动态分类、度量和预测商业周期，以及量化媒体的政治倾向等问题。在金融学研究领域，文本大数据主要用于度量投资者和媒体关注度、不同市场参与主体的情绪、基于新闻的隐含波动率以及投资者的意见分歧等指标。我们对这些应用研究的数据来源、处理方法和分析结果进行了全面的梳理。在此基础上，本文还讨论了基于文本大数据的实证分析的新特征以及未来可能的研究趋势。

关键词：文本大数据，机器学习，深度学习，不确定性，投资者情绪

JEL 分类号: C42, G12, G14

文本大数据分析在经济学和金融学中的应用：

一个文献综述[1]

沈艳 （Yan Shen）

单位：北京大学国家发展研究院，北京大学数字金融研究中心

National School of Development, Peking University

Institute of Digital Finance, Peking University

联系电话：

邮箱：yshen@nsd. pku. edu. cn

陈赟 （Yun Chen） （通讯作者）

单位：北京大学国家发展研究院

National School of Development, Peking University

通讯地址：北京大学国家发展研究院, 100871

联系电话：

邮箱：yunchen@pku.edu.cn,

黄卓 （Zhuo Huang）

单位：北京大学国家发展研究院，北京大学数字金融研究中心

National School of Development, Peking University

Institute of Digital Finance, Peking University

联系电话：

邮箱：zhuohuang@nsd.pku.edu.cn

[1] 本研究受国家自然科学基金面上项目（编号 71671004）和国家社会科学基金重大项目（编号 18ZDA091）资助。

文本大数据分析在经济学和金融学中的应用：

一个文献综述

本文对文本大数据分析在经济学和金融学中应用的文献进行综述。文本大数据具有数据来源多样化、数据量增长快和时频高的特征，能够为经济学和金融学研究提供新的信息和独特的分析视角，但是处理文本大数据存在许多挑战。我们梳理了文本大数据的信息提取步骤，总结了常用的词典法、机器学习方法和深度学习方法的实现原理和技术特点。在经济学研究领域中，文本大数据主要用于测度经济和政策不确定性、基于文本的行业动态分类、度量和预测商业周期，以及量化媒体的政治倾向等问题。在金融学研究领域，文本大数据主要用于度量投资者和媒体关注度、不同市场参与主体的情绪、基于新闻的隐含波动率以及投资者的意见分歧等指标。我们对这些应用研究的数据来源、处理方法和分析结果进行了全面的梳理。在此基础上，本文还讨论了基于文本大数据的实证分析的新特征以及未来可能的研究趋势。

关键词：文本大数据，机器学习，深度学习，不确定性，投资者情绪
JEL 分类号: C42, G12, G14

A Literature Review of Textual Analysis in Economic and Financial Research

Abstract: This paper conducts a literature review of textual analysis in economic and financial studies. Textual data exhibit the characteristics of diverse data source, rapidly growing data volume and high frequency. While textual data bring new information and perspectives in economic and financial research, many challenges remain in effectively dealing with textual data. We summarize the procedures of extracting information from textual data and discuss mechanism and features of the popular methods such as dictionary-based approach, machine learning and deep learning approaches. In economic research, textual data have been used in measuring economic policy uncertainty, text-based network industry classification, monitoring and predicting business cycle and quantifying media slant. In financial research, textual data have been used in measuring investor attention and sentiment, news implied VIX and investor disagreement. We review the data sources, approaches and empirical results of these studies. Finally, we summarize the features of using textual analysis in empirical studies and point out future research directions in this field.

Keywords: Textual analysis, Machine learning, Deep learning, Uncertainty, Investor sentiment

JEL Classification: C42, G12, G14

一、 引言

得益于互联网的快速发展和计算机技术的进步，文本大数据在经济学和金融学领域的应用方兴未艾。在经济学领域，文本大数据被用于刻画经济政策不确定性（Baker et al, 2016）、对行业进行动态分类（Hoberg and Phillips, 2015）、度量和预测经济周期（Thorsrud，2018；Shapiro et al., 2018），和度量媒体政治倾向及新闻需求（Gentzkow and Shapiro, 2010）等问题。金融学中，文本数据被用于刻画关注度（如 Antweiler and Frank, 2004；Fang and Peress，2009; Garcia, 2013）、情绪或语调（如 Tetlock 2007； Li 2010；Da et al. 2011；Loughran and McDonald 2011; Jegadeesh and Wu 2013；Kim and Kim 2014；Tsukioka et al. 2018）、新闻隐含波动率（Manela and Moreira, 2017）和意见分歧（Antweiler and Frank, 2004; Hillert et al. 2018）等方面。非结构化文本大数据的运用拓宽了经济和金融领域的实证研究，但也带来了新挑战。

作为新数据源，文本大数据至少有三个特征。一是数据来源多样化。相对于主要由政府和机构主导收集的传统数据，文本大数据的发布主体有个人（如投资者、消费者）、企业、媒体、机构和政府相关职能部门等；其具体形式丰富多样，如 Twitter、微博、论坛帖子、消费者对产品的评价、微信公众号、上市公司年报、电话录音文稿、招聘广告、公司年报、季报、公告、IPO 招股说明书、分析师研究报告、会议纪要、有影响力的政治、经济、金融领域人物的演讲、央行等政府机构定期和不定期发布的各类信息等。二是数据体量呈几何级增长。囿于数据收集成本，传统数据往往需要借助纸质媒介，体量较小。随着文本信息从纸质媒介向以互联网为媒介的方式转移，文本数据收集和传输成本大幅降低，为计算机领域的自然语言处理方法（Natural Language Processing，NLP）提供了应用场景。三是时频高。传统数据需要经过系统性的组织和安排来收集，常用的经济和金融领域数据多为年度、季度、月度、周度数据，频率更高的数据可得性不足，不足以满足对经济和金融领域高频数据分析的应用需要。而文本大数据的频率可以高达秒级（如网民在网络平台上发布的消息和观点的时间颗粒度），这为高频研究提供了数据基础。

文本大数据为经典研究问题提供了新视角（Gentzkow et al.，2017），也可用于研究新的问题。例如，投资者情绪如何影响资产定价是经典问题，投资者情绪的度量是实证研究的关键步骤。传统度量方法包括选择市场变量作为投资者情绪代理变量的市场变量法和采用调查问卷收集到的答案来度量情绪的调查法。Baker and Wurgler （2006）对六个市场变量采用主成分分析法构建的情绪指数采用了市场变量法，而密歇根大学消费者信心指数则是调查法主要代表。由于市场变量法获得的指数更可能是关于情绪的均衡结果（Qiu and Welch，2004）、因此不只包含情绪(Sibley et al. 2016)，而调查法频率低、成本高、受访者答案未必是其真实意图表述，现有投资者情绪度量手段均有缺陷。通过收集反映投资者情绪的言论形成的文本数据（如论坛帖子、微博）提供了直接度量情绪的新渠道。又如，经济不确定性是影响宏观经济周期的重要因素（Bachmann et al.，2013；Baker et al.,2016），也会对金融市场产生影响（Bali et al. 2017），除了采用市场变量近似不确定性的方法外（Jurado et al.，2015），采用新闻文本数据构造的经济政策不确定性指数不仅时频高，也为各国各地区采用统一标准度量不确定性提供了可能（Baker et al.，2016）。再如，Manela and Moreira（2017）利用 1890-2009 年间华尔街日报的新闻构造的新闻隐含波动率指数（News Implied Volatility），不仅为理解波动率提供了新渠道，还近似出尚不存在 VIX 指标的历史金融市场的风险状况。

将文本大数据应用于经济学和金融学研究的核心挑战在于如何准确、有效率地从文本中提取出需要的信息，并考察其对相应问题的解释或预测能力。如图一所示，令 Ψ 代表采用的原始文本库，Y 代表要解释或者预测的经济或金融现象，要考察 Ψ 对 Y 的解释能力，

需要经过三个步骤。第一，将文本库 Ψ 内所有文本转化数据矩阵 Λ；第二，通过计量或者统计方法 F，将 Λ 转换成目标信息序列 V，如关注度、情绪、不确定性等指数；第三，用提取出的 V 来解释或预测 Y。

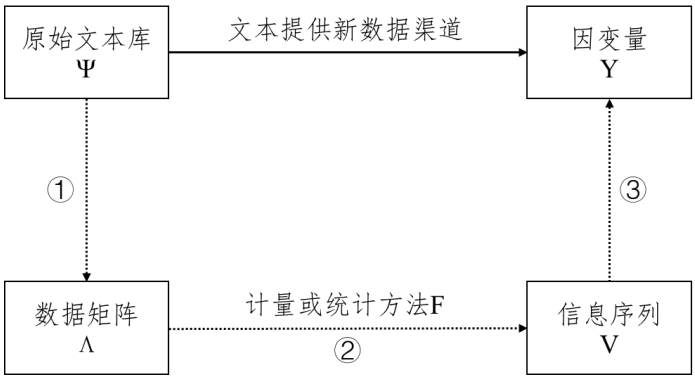

图 1 文本信息提取步骤

目前对实现上述三步转换需要解决的问题和相应实际应用，现有中文文献缺乏较为全面和系统的梳理。本文有两个主要目标，一是介绍从原始文本库 Ψ 到解释或预测 Y 的过程中，不同步骤面临的主要挑战、解决方法及其特点。由于第三步（利用结构化数据来完成解释或预测 Y 的工作）是计量经济学和统计学的研究重点，本文侧重介绍前面两步涉及到的方法。第二个目标是简要梳理国内外运用文本数据在经济和金融领域的应用，并探讨未来的研究方向。我们在第二部分介绍文本信息提取方法，即如何从原始文本库出发，提取出需要的信息序列。在第三部分本文梳理经济和金融各领域文本数据的应用，第四部分为结论和展望。

二、 文本大数据信息提取方法

从中文原始文本中提取出研究者需要的信息的过程，需要让计算机学习用类似人的思维模式来分析和处理语言。本部分着重介绍将非结构化的帖子、文章、演讲词、会议记录等原始文本库 Ψ 作为输入，通过一定的转换方法得到结构化数据矩阵 Λ；再以数据矩阵 Λ 为输入，利用统计或者计量模型，输出目标信息序列 V 涉及的方法。

（一）原始文本库到数据矩阵的结构化转换

从形式上看，一个中文文本是由汉字（包括标点符号等）组成的一个字符串。如果将文本从大到小分解，可能得到篇、章、节、段、句子、词组、词和字。自然语言理解中的主要困难和障碍，是同一个字（词）的含义在不同的场景或语境下有变化；同时由于文字的丰富多样性，在转换为数据矩阵后往往需要处理高维稀疏矩阵相关的问题。在本小节着重介绍确定文本数据基础单位的分词技术和词嵌入（word embedding）技术，即将词转换为向量的方法。经过上述转换，非结构化文本可用矩阵形式表示，其中每一行记录同一个体的不同属性信息，而同一列数据记录不同个体的同一属性相关资料。

1. 分词技术

在英文环境下，单词被空格分隔开，因此单词就实现了分词。实证运用中也会将单个词语扩展成长度为 *n* 的词组，即 *n* 元词组(*n-gram*)。例如 Gentzkow and Shapiro（2010）发

现，在分析不同党派的演讲内容时，词组比单词包含更多的信息，也更能反应政党的语言色彩。由于 n 的值越大，总词组的数量就越多，表示文本的矩阵或向量的维数将呈几何级增长，因此常用的 *n-gram* 模型中 n 取值一般为 1、2 和 3。

由于中文中汉字为连续序列，分析文本就需要按照一定的规范将汉字序列切分成词或词组，即中文分词。根据分割原理，可将现有的分词方法归纳为基于字符串匹配、基于理解和基于统计这三类。字符串匹配法将待分析的汉字串与前定的词典词条匹配，若某个字符串可在词典中找到，则记为识别出一个词。该方法的好处是简便快速，但忽略歧义问题。基于理解的分词方法则在分词的同时进行句法、语义分析，以改进对歧义词的处理。基于统计的分词方法则先用机器学习模型学习已经切分好的词语的规律，进而实现对未知文本的切分，常用方法包括最大概率分词法和最大熵分词法等。

目前经济和金融文献中用到的中文分词方法往往能结合上述三种方法，如自然语言与信息检索共享平台 NLPIR[2]（汪昌云和武佳薇，2015），中科院汉语词法分析系统（段江娇等，2017），Python 软件包“jieba”（王靖一和黄益平，2018；Chen 等，2018）等。需要注意的是，由于一些特定领域的文本包含一些对信息提取比较重要的专有词语（如上市公司名称、金融术语等），因此常常需要根据研究问题拓展现有词典，以提高软件识别和分割词语的准确度。

2. 词转换为向量的技术

完成分词之后需要完成的是如何将文本进一步转化为数字化矩阵。如果将一篇文本视作从所有词语库中挑选若干词形成的组合，这一转换的主要挑战往往是如何对由词语构成的高维矩阵实现降维的问题。要理解这一点首先需要介绍独热表示法（One-hot representation）。

（1） 独热表示法

独热表示法最早的应用是在自然语言处理和信息检索领域。在金融领域，Manela and Moreira （2017）使用该方法，根据华尔街日报新闻的标题和摘要中全部词语出现的频率来提取新闻数据的特征。

独热法的特点是忽略语法和语序等要素，将文本数据看作是若干独立词汇的集合。首先，根据文本中出现的全部词语构建一个词表[3]，并将每个词按顺序编号$1，2，3…，N$。然后，将词语j用一个N维向量w_j来表示，该向量的第j个位置的元素为 1 其余均为 0。在每一个词都转换为一个向量后，通过加总所有词的向量，文本 t 就可以转化为 $1\times N$的向量W_t，其中$w_{tj}, j = 1,\dots,N$,是第j个词语在文本 t 中出现的频率。若一共有$t = 1,\dots,T$个文本，采用独热法之后，原始文本库 Ψ 就可以转化为 $T\times N$的数字矩阵。

例如，原始文本库 Ψ 由两条帖子组成。第一条的内容是“明天涨停。后天涨停没戏。”，第二条是“玛丽有个小绵羊。”分词后得 “明天、涨停、后天、没戏、玛丽、有、个、小、绵羊”九个不同词语，即$N = 9$。用独热法则“明天”用向量[1,0,0,0,0,0,0,0,0]表示，“涨停”为[0,1,0,0, 0,0,0,0,0]，以此类推。于是第一个帖子可用向量[1,2,1,1,0,0,0,0,0]表示,第二个帖子即[0,0,0,0,1,1,1,1,1]。

上述步骤显示，独热法操作简单；但数据量大时转换后的矩阵往往是高维稀疏数据矩阵。这是由词向量维数由词语数量决定、并且大部分词语出现频率低，因此文本对应的向量中绝大部分值为零的特征决定的。另外，独热法可能因忽略上下文结构而会产生歧义。例如上例中第一个帖子转换成的向量也可以是 “明天涨停没戏。后天涨停。”的转换结果，这和原文的含义产生了偏差。

[2] http://www.nlpir.org/。

[3] 通常是去掉“的、地、得、和”等停用词和标点符号后得到的全部词语。

要解决文本数据是高维稀疏矩阵的问题有两种策略，一是采取多种措施对数字化文本矩阵实现降维，Gentzkow 等(2018)对相应降维方法已经做了系统总结。另一个思路则是采用词语嵌入技术（Word Embedding），直接在词语转换成数字化矩阵时就将词语转化为低维向量。

（2）词嵌入技术

词嵌入技术是指把一个维数为所有词的数量的高维空间“嵌入”到一个维数低得多的连续向量空间中涉及的模型和技术，即，其中e_j表示第j个词通过嵌入矩阵 E(embedding matrix)映射到实数域上的词向量[4]。由于该向量的每个元素值可以是连续值而不只是 0 或者 1，e_j的维度N_e可以远低于N。

独热法可以看做是最简单的词嵌入方法，即$e_j=E\cdot W_j=W_j$。常用词嵌入算法包括 Mikolov et al.（2013）提出的词向量 Word2Vec 和 Pennington 等人开发的 GloVe（Global vectors for Word Representation），其中 Word2Vec 的应用似更为广泛。Word2Vec 的主要思想是先用向量代表各个词，然后通过神经网络模型，在大量的文本语料数据上来学习这些向量的参数。训练后的模型不仅可以将每个词语映射到一个低维的空间上（通常为 100-1000 维），每个维数上的取值为连续值；并且根据不同词语的向量距离可以度量词语间的相似程度，也解决了独热法下不同词语相互独立的问题。

Word2Vec 技术在计算语言学等领域得到了广泛的应用，并且和其他的统计模型结合在一起进行文本分析时发现具有很好的表现，但在经济金融领域的应用相对较少（如 Gentzkow et al.，2018）。近年来，该方法也逐渐得到重视。例如，王靖一和黄益平（2018）利用该技术来拓展了金融科技情绪词典。Chen et al.（2018）对比了独热法和 Word2Vec 两种词向量表示方法，发现 Word2vec 的分类准确性可高达 82%，远高于独热法。

（二）数据矩阵的信息提取

现有经济和金融领域文本相关分析的问题大致可分为两类，一是区分文本显示的投资者情绪正负、新闻或者文件语调正负、报纸属于左派还是右派、以及行业分类等聚类问题，二是对情绪、不确定性、恐慌程度、意见分歧程度的度量以及相应的回归问题。根据事先是否有存在有标签的训练数据，这些问题可以采用有监督学习或无监督学习这两类方法来分析。其中，无监督学习的主要方法包括词典法和主题分类模型等，而支持向量机等机器学习经典方法和深度学习方法近年在经济和金融领域的运用更多属于有监督学习。

1. 无监督学习方法

（1） 词典法

词典法是一种传统的文本大数据分析方法。该方法从预先设定的词典出发，通过统计文本数据中不同类别词语出现的次数，结合不同的加权方法来提取文本信息。在经济金融领域中，词典法得到广泛运用（如 Tetlock，2007；Tetlock et al，2008；Laughran and McDonald，2011；Garcia，2013；Da et.al, 2014; Renault，2017；Chen et al.，2018 等）。

使用词典法的一个关键环节是选择或构建合适的词典，这里词典包括了特定词典，也包括作者构造的特定词语或词组的集合。文献中常用的英文特定词典包括 Harvard IV-4 心理社会词典[5]、Henry 词典、Diction 词典[6]和 Loughran and McDonald 词典[7]。早期文本情绪构造

[4] 吴恩达《深度学习》，https://blog.csdn.net/u013507678/article/details/80686382。

[5] http://www.wjh.harvard.edu/~inquirer/homecat.htm。

[6] https://www.dictionsoftware.com/。

[7] https://sraf.nd.edu/textual-analysis/resources/。

多使用 Harvard IV-4 词典（Tetlock，2007；Tetlock et al.，2008；Jegadeesh and Wu，2013），它包含心理和社会学常涉及的 1045 个正面词语和 1160 个负面词语[8]，但并非为金融领域文本专门创建。Henry 词典是专门为金融文本构建的词典（Henry, 2008），Price et al. （2012）认为 Henry 词典比 Harvard IV-4 词典更准确地度量了上市公司盈利披露电话会议文字稿中的语调，但它包含的负面词汇较少。Diction 词典包含 686 个正面词汇和 920 个负面词汇，主要应用于会计领域（Rogers et al.，2011；Davis et al.，2012）。Loughran and McDonald（简称 LM）词典由 Loughran and McDonald（2011）从上市公司的 10-K 文件中人工收集并整理构造出来，他们的实证结果表明 LM 词典在度量文本情绪方面比 Harvard IV-4 词典和 Diction 词典的效果更好，因此目前用词典法分析金融、会计领域文本情绪时多采用 LM 词典（Garcia，2013； Huang et al.， 2014； Loughran and McDonald，2014；Solomon et al.，2014 等）。

在特定词典外，使用作者自行构造的词或者词组的代表性研究包括 Da et al.（2014），Hoberg and Phillips (2015)，和 Baker et al.（2016）。Da et al.（2014）选取了 118 个与经济相关的词语（词组），利用这些词语在谷歌搜索中的搜索频次，构建了用来度量投资者情绪的 FEARS（Financial and Economic Attitude Revealed by Search）指数。Hoberg and Phillips (2015) 采用 10-K 文件中对上市公司产品的描述，将不同上市公司作了基于文本的行业分类。Baker et al.（2016）则选取和经济、政策、不确定三个类别相关的一些词语，通过统计同时包含这些词语的新闻的比例，构建了经济政策不确定性（Economic Policy Uncertainty ，EPU）指数。

在中文语境下使用词典法，需要注意的是直接翻译的英文词典可能并不适用。Chen et al.（2018）随机抽取 2008 到 2018 年间某股票论坛四万条帖子，人工挑取其中正、负面词语，构建了适用于中国股吧论坛的金融情绪词典。他们发现，与直接使用翻译的 LM 词典相比，该词典能将情绪分类准确率提高 30%。中文相关文献有作者就具体问题构建的中文词典。例如，汪昌云和武佳薇（2015）手动整理新闻报道，结合《现代汉语词典》、《最新汉英经济金融常用术语使用手册》、LM 词典中文版以及知网-中文信息结构库等词库，构建了中国财经媒体领域的正负面词库。王靖一和黄益平（2018）根据和讯网上的新闻，构建了适用于金融科技领域的情感词词典等。

在确定词典外，用词典法分析文本情绪是另一个要处理的问题是如何确定词语权重。Jegadeesh and Wu（2013）指出，选择合适的加权方法有时比构建完备且精确的词典更重要。常用的加权方法有等权重、词频-逆文档（TF-IDF）加权，和对应变量加权这三种。顾名思义，等权重法假定文本中每个词语的重要程度相同。TF-IDF 加权方法则同时考虑词语在文本中出现的次数（频率）和多少文档包含该词语这两个维度，对在文本中频繁出现但并没有实际含义的词语赋予较少的权重、而给予有重要含义但出现次数较少的词语较大权重。对应变量加权是指借用文本中词语与对应变量（市场收益率、波动率指数等）的关系来确定词语的权重。

不同权重法各有千秋。等权重法因简便易行而广为使用（如 Hoberg and Phillips，2015；Box, 2018），不过 Loughran 和 McDonald（2011）发现在 10-K 文本下，TF-IDF 加权法比等权重法更可降低词语分类错误，可以实现更为有效的信息提取。Chen et al.（2018）的研究结果也表明，在中文语境下使用 TF-IDF 能够比等权重方法得到更准确的情绪分类。对应变量加权法的优点是能在一定程度上避免权重选择的主观性，其效果也不依赖于词典是否完整，因此文本分析结果可能比人为主观设定权重更为准确。例如，Jegadeesh and Wu（2013）发现利用 10-K 文本数据和异常收益率计算出的词语权重来计算 IPO 招股说明书的语调得分，能够用来解释 IPO 折价现象。Renault（2017）根据网络论坛 StockTwits 上带有标签（看

[8] http://www.wjh.harvard.edu/~inquirer/homecat.htm。

涨和看跌）的帖子，根据这两类文档出现的频率，为不同词语正负情感的强弱程度作加权。Da et al.（2014）构建 FEARS 指数和 Manela and Moreira（2017）构建 NVIX 指数均采用了类似的思想。当然，对应权重加权的弱点是选择作为词语权重的对应变量具有主观性，对一项研究恰当的词语权重法未必能适用于另一项研究。

总体而言，只要运用得当，词典法从文本中提取信息的能力较强，这种优势对于短文和对于词语间逻辑关系较弱的应用更为明显。例如，Renault（2017）发现词典法和机器学习方法在识别论坛情绪准确率方面不相上下。Chen et.al（2018）对中文论坛帖子数据的分析也有类似的结论。由于词典法对于有监督学习和无监督学习均有应用场景，常常可以作为文本大数据分析的一种基准方法。

（2） 主题分类模型

在经济和金融领域的一个应用需求是在没有事先标注集的情况下，对文本按主题做分类。由于一篇文本的主题可能有多个，这类分类问题不同于按照事先标注集、将一篇文本仅归入一类的应用。主题分类问题的代表模型是由 Blei et al.（2003）提出的隐含狄利克雷分配（Latent Dirichlet Allocation，LDA）模型，它是一种概率主题模型。

LDA 模型假定全部文档M中存在K个主题，每个文档m包含N_m个词语，并且每个词都是由其中的一个主题生成。主题服从一个多项式分布$\boldsymbol{\theta}_m$，而每个主题k与词汇表中的V个单词的一个多项式分布$\boldsymbol{\varphi}_k$相对应，并且假定分布$\boldsymbol{\theta}_m$和分布$\boldsymbol{\varphi}_k$具有共轭的狄利克雷分布，该共轭的狄利克雷分布的超参为α和β。这样，通过预设文档中的主题个数，LDA 模型可以将每篇文档的主题以概率分布的形式给出，其中每个主题对应一类词语分布，根据词语分布可以挑选出一些关键词对该主题进行描述。

LDA 模型假定文档的生成过程如图二所示：（1）从狄利克雷分布α中抽样得到文档m的主题多项式分布$\boldsymbol{\theta}_m$，从狄利克雷分布β抽样得到主题k的词语多项式分布$\boldsymbol{\varphi}_k, k = 1, \dots, K$；（2）从主题多项式分布$\boldsymbol{\theta}_m$中抽样得到文档$m$的第$n$个词的主题$Z_{m,n}$；（3）从主题$Z_{m,n}$对应的词语分布$\boldsymbol{\varphi}_{Z_{m,n}}$抽取词语$W_{m,n}$；（4）重复上述步骤$N_m$次。因此，所有已知的和隐藏的变量的联合分布可以表示为：

$$P(\boldsymbol{W}_m, \boldsymbol{Z}_m, \boldsymbol{\theta}_m, \boldsymbol{\Phi}; \alpha, \beta) = \prod_{n=1}^{N_m} P(\boldsymbol{\theta}_m; \alpha)\, P\left(Z_{m,n} \middle| \boldsymbol{\theta}_m\right) P(\boldsymbol{\Phi}; \beta) P\left(W_{m,n} \middle| \boldsymbol{\varphi}_{Z_{m,n}}\right),$$

其中$\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_k\}_{k=1}^{K}$，模型中唯一可观测的变量是词语$W_{m,n}$。实际应用中可以通过 Gibss 抽样方法来估计 LDA 模型的参数，从而得到每篇文档的主题分布$\boldsymbol{\theta}_m$和每个主题对应的词语分布$\boldsymbol{\varphi}_k$。

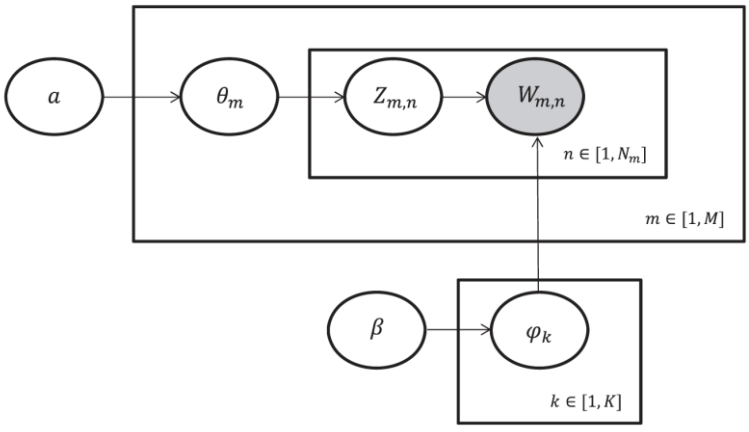

图 2 LDA 模型图示法。图片来自 Thorsrud（2018）

LDA 的一个局限性是需要人为地给出一个主题数量，而主题数量的选择会影响主题的生成和文档的归类。选择文档主题个数K的方法有：根据主题个数计算得到复杂度得分（perplexity score；Blei et al.，2003），交叉验证法（Airoldi et al.，2010），预设一些初始值、再根据主题的解释能力来调整主题个数（Gentzkow et al.，2018）等。LDA 模型的一个拓展是 Teh et al.(2006)提出的层次狄利克雷过程（Hierarchical Dirichlet Processes）。该方法不需要事先设定 K，而是将主题个数作为未知的模型参数并结合贝叶斯非参技术来估计。LDA 的另一个局限性是忽略了主题分布随时间可能存在的演进变化，相对应的拓展是 Blei 和 Laffery(2006)的动态主题模型（Dynamic Topic Models）。对这些拓展的细节本文不复赘述。

近年来主题分类模型在经济和金融领域逐渐得到运用。例如，Thorsrud（2018）使用 LDA 模型从新闻数据中提取出了 80 个主题并估计日度频率的新闻即时经济周期指数；Wang et al.（2018）使用 LDA 和 HDP 从近两千万新闻中分离出金融科技主题，并且构建了金融科技情绪指数。

2. 有监督学习方法

（1） 经典的有监督机器学习方法

经典机器学习方法包括朴素贝叶斯、支持向量机、决策树、K 近邻算法、adaboost、k 邻近算法，最大熵法等。在金融领域的文本分析中，较为常用的传统机器学习方法包括朴素贝叶斯（Naïve Bayes）和支持向量机（Support Vector Machine，SVM）。

朴素贝叶斯算法（Murphy，2012）是一种基于贝叶斯理论的有监督学习算法。在处理文本分类问题时常见步骤如下。首先根据训练集学习文本中词语与所属类别的关系，得到朴素贝叶斯分类器的先验分布（即文本属于不同类别的先验概率），以及条件概率分布（即给定分类类别下某个词语出现的概率）。其次，使用这些概率，根据文本中的词语特征，结合贝叶斯条件概率公式，计算该文档属于不同类别的条件概率。最后，按照最大后验假设将文本分类为具有最大后验概率的一类。

Antweiler and Frank（2004）较早采用朴素贝叶斯方法研究文本情绪。他们挑选了 1000 条雅虎财经上的帖子，人工将其分类为买入、卖出、持有。接着利用这些人工标注帖训练朴素贝叶算法，并将其应用到剩余的未分类帖子上，最后根据买入、卖出帖子的数量构建了 Bullishness 指数用于度量文本情绪。此后，在经济金融领域不少文献将朴素贝叶斯算法应用到不同类型的文本上，如雅虎财经上的股票讨论帖（Das and Chen，2007；Kim and Kim，2014）、公司年报文件（Li，2010；Jegadeesh and Wu，2013）、分析师报告（Huang et al.，2014）、新闻报纸（Buehlmaier and Zechner，2016），和网络论坛发帖（段江娇等，2017）。

朴素贝叶斯算法假定文本中的特征互相独立，即不同词语不存在相互依赖关系。支持向量机（Vapnik，1996）是一种容许词语间存在相互依赖的有监督学习算法，既可以用于分类也可以用于回归分析。其基本原理是，首先将每个文本投射为N_e空间的一个点，通过寻找到一个超平面，将这些点按照其对应的标签（如正、负情绪等）进行分割，使得每个类别的点到这个超平面的最近距离最大化。使用支持向量机进行分类和回归分析前的步骤包括，首先采用独热法或者 Word2vec 等方法将文本转换为向量，然后根据训练集学习文本向量与所属类别的关系，再对将根据训练集得到的模型做交叉验证（cross-validation），最后将训练出的最优模型用于预测所有文本的分类。

SVM 相关应用主要出现在金融领域。如 Manela and Moreira（2017）使用独热法，将 1890-2009 年间华尔街日报头版新闻向量化；再使用支持向量回归法提取新闻隐含波动率指数。Tsukioka et al.（2018）使用 SVM 方法度量日文环境下的股票论坛投资者情绪。Chen et al.（2018）使用 SVM 对中国网络论坛帖子进行情感分类。

此外，k 邻近算法（杨晓兰等，2016），最大熵法（Renault，2017）等也有运用，其表现和朴素贝叶斯、支持向量机方法相近。

（2） 深度学习法

文本分析中，SVM 等分类器虽然可以处理一定的非线性，但作为线性分类器，这类方法往往只能将输入数据切分为非常简单的区域，也容易导致过拟合等问题（Gentzkow et al. 2018）。随着大数据可得性的增加、人工智能软硬件技术的发展，深度学习方法在自然语言处理领域的强大功能逐渐显现。作为机器学习的分支，深度学习试图通过模仿人脑的神经网络，使用多重非线性变换构成的多个处理层对数据进行高层抽象，以实现分类等目标。这类方法可用于有监督和无监督学习，但目前尚未在经济领域有广泛运用，在金融领域的运用主要是有监督学习（Chen et al., 2018）。

神经网络（Neural Network；Bishop，1995）是基于模仿人脑的神经网络来实现人工智能的机器学习模型，包含输入层、隐藏层、输出层等结构，可用于处理文本分类问题，其原理是输入层的特征向量通过隐含层的变换到达输出层，在输出层得到分类结果，通常使用反向传播算法对神经网络模型进行训练。

深度学习常用模型包括深度神经网络（Deep neural nets，DNN）、卷积神经网络（Convolutional Neural Network，CNN），和循环神经网络（Recurrent Neural Network，RNN）等。作为神经网络模型的拓展，DNN（Hinton and Salakhutdinov，2006； LeCun et al.，2015）可以通过增加网络层数、减少每层网络节点数，以及使用不同的传输函数克服训练过程中的梯度消失现象等方法，可用于处理文本分类、翻译、语义分析等复杂的自然语言处理任务。针对 DNN 参数数量巨大，没有考虑数据的固有局部特征等缺陷， Kim （2014）提出了 CNN 方法。CNN 模型进行文本分类时，不仅限制了参数个数，还通过考虑词语在文本中的上下结构来挖掘文本内的局部结构。Kim（2014）发现，将 CNN 和 Word2vec 嵌套在一起使用可以在对文本进行分类时达到非常高的准确率。第三种深度学习方法 RNN（Elman，1990）处理文本分类问题的思想是借用 RNN 模型的递归结构来捕捉上下文信息。深度学习的常用模型还在不断拓展中，如 CNN 和 RNN 模型组合在一起的 RCNN 模型（Lai et al.，2015）等。这些深度神经网络模型的好处是可以提供非线性分类，但代价是模型待训练参数很多，通常一层结构的 CNN 模型就需要训练数万的参数，因此训练样本足够大是模型效果的基本保障。

目前在经济金融领域使用深度神经网络模型提取文本信息的文献较少。Chen et al.（2018）首次采用 CNN 来计算中国散户投资者情绪，并比较了 CNN 与 SVM 等模型的预测效果。他们的研究发现，在采用四万条训练数据集的情况下，训练出的 CNN 模型的预测准确性与 SVM 大致相当，但是在分类中 CNN 模型的分类更为果断；随着训练数据集的增大，CNN 的优势可能会进一步显现。

无论是采用经典机器学习方法还是新兴的深度学习法，有监督训练都需要两个要素：高质量的标注数据作为训练集和明确的模型选择标准。由于训练集质量会直接影响最终信息提取效果，做相关研究应事先评估构建标注数据需要耗费的成本。在模型选择标准方面，理想模型不仅要能避免样本内过拟合，也要有较好的样本外表现。通常需要采用交叉验证的方法来评估模型：首先将标注集按照一定的比例随机分为训练集、验证集和测试集；再在训练集上训练模型，根据其在验证集上的表现来调整模型参数，再将模型应用到测试集上计算准确率以挑选最优模型。

综上所述，选择文本数据信息提取方法需综合考虑文本数据的来源、语言环境、内容长短以及需提取信息的特征等因素，同时评估各类方法的成本和收益。在条件允许的情况下，可同时考虑简单方法和复杂方法，通过分析比较两类方法的差异来提高信息提取的准确性。当然，使用复杂方法时需要保证这些方法的透明性和可复制性（Loughran and McDonald，

2016）。最后还要注意的是，数据的结构化转换和文本数据信息提取这两步的执行顺序需要依靠具体问题来决定，有时需要反复尝试才能找到最佳方案。

三、 文本大数据分析在经济学和金融学的应用

目前，文本大数据分析在经济学和金融学中的应用日渐广泛。经济学应用主要体现在刻画经济政策不确定性、预测通货膨胀和失业、对行业进行动态分类，和度量媒体政治倾向及新闻需求等问题中。在金融学中的应用主要包括在刻画关注度、情绪、隐含波动率、和意见不一致性等四个方面的指标。

（一） 在经济学中的应用

1. 经济政策不确定性指数

经济政策不确定性是经济中的个体对未来政策的变动和当前政府政策影响的不确定性程度（Gentzkow et al.，2018）。传统方法从市场变量出发来度量经济不确定性[9]，这些度量方法存在市场变量时间跨度短、频率低，不同国家指标不具可比性等弱点。而各国主流新闻媒体的新闻文本历史跨度长、频率高，因此 Baker et.al（2016）另辟蹊径，采用新闻文本数据来度量经济政策不确定性的 EPU 指数。

从构建方法看，EPU 指数属于作者自行选择词或词组的词典法。首先作者收集美国 10 家主流新闻媒体从 1985 年以来的新闻数据，使用机器自动统计各媒体新闻中同时包含经济（economic /economics）、不确定（uncertain/ uncertainty）和政策[10]三类词语的月度文章数量。为控制新闻数量的时间趋势，作者对上述文章数量作标准化处理，再对这十个标准化的序列按月作平均，最后将序列转换为均值 100 的指数。

在从三个角度验证了 EPU 指数能较好度量经济政策不确定性后[11]， Baker et.al（2016）对 EPU 指数还做了一系列拓展：（1）构建 11 个主要经济体国家月度频率的 EPU 指数；（2）构建货币政策、财政政策、国防等 11 个政策分类 EPU 子指数；（3）构造英国和美国 1900-2011 日度 EPU 指数。

研究 EPU 与其他经济变量间的关系是近期文献的热点之一。如 EPU 对企业层面的股价波动率、投资率、就业增长率的影响，对市场加总层面的投资、产出、就业的影响（Baker et.al，2016）；对公司投资的影响（Gulen and Ion，2016）；对股市波动率的影响 Paster and Veronesi，2013）；对市场超额收益率的预测（Brogaard and Detzel，2014）等。国内也有不少文献使用 Baker et al.（2016）提供的中国市场 EPU 指数来研究经济政策不确定性对微观企业经营活动和决策行为的影响，如经济政策不确定性如何影响公司现金持有水平（王红建等，2014）、企业投资行为（李凤羽和杨墨竹，2015；饶品贵等，2017）、企业创新（顾夏铭等，2018）、国有企业和非国有企业的杠杆率（纪洋等，2018）、分析师盈余预测修正（陈胜蓝和李占婷，2017）、企业金融化趋势（彭俞超等，2018）、企业资本结构动态调整（顾研和周强龙，2018）和公司提供的商业信用（陈胜蓝和刘晓玲，2018）等。

---

[9]如反应对权益资产未来收益不确定性的市场波动率指数 VIX 或 VXO、衡量投资者的风险厌恶程度的方差风险溢价（Bali and Zhou，2016）、横截面分散程度（上市公司股票收益率、企业利润增长率、分析师 GDP 预测、不同行业的全要素生产率等变量的标准差）、基于大量经济变量中不可预测部分条件波动率的 Jurado et al.（2015）不确定性指数等。

[10]政策类词语为 Congress，deficit，Federal Reserve，legislation，regulation，White House，从 1985-2012 年新闻中出现频率最高的 15 个词语中选出。

[11] 这三个角度是（1）EPU 指数和重大经济政治事件相吻合；（2）采用人工阅读 12000 多份报纸构建的指数和使用机器构建的两种 EPU 指数相关系数高达 0.86；（3）EPU 指数与其他常用的不确定性度量指标有很高相关系数、特征相似。

当然，基于英文南华早报度量的中国市场 EPU 指数也存在一定的局限性（饶品贵等，2017）。目前度量中国经济不确定性的指标除了 EPU 外，还有 Huang et al.（2018）参照 Jurado et al.（2015）的方法构建的经济不确定性指数。目前从中文文本数据出发度量中国经济不确定性的研究尚属空白。

2. 行业分类

产业组织中的一个核心问题是定义行业边界和行业竞争力，传统的行业分类通常较为固定（如标准行业分类 SIC 和北美产业分类体系 NAICS 等）。Hoberg and Phillips（2016）根据上市公司年报中对企业产品描述的内容提出了一种新的行业分类方法，即文本网络行业分类法（text-based network industry classification）。首先统计全部文档中包含不同词语的个数 W，并构建对应的词表；接着采用独热法将每个企业的产品描述文档转换为长度为 W 的向量，即如果文档中包含某词语，则对应位置的元素取值为 1，否则为 0。然后，计算不同文档向量的余弦相似性作为不同公司产品相似度的度量。最后根据这些相似度得分，结合聚类算法，可以将不同公司分组到不同的行业，最终得到 300 个行业分类（与 SIC 和 NAICS 中的行业数量一致）。

基于这种时变的行业分类标准，他们检验了公司如何对产品市场的外部和内部变化做出反应，并评估在军事和软件行业等大的外生行业冲击下，企业对产品生产调整的反应程度。他们发现外生行业冲击会对相似企业数量、产品差异性、企业生产的产品种类等产生重大影响。除了用于研究外生冲击对行业内竞争和企业产品生产决策的影响外，作者还认为上述分类可用于解释行业内盈利能力、销售增长率、市场风险等不同特征的差异性。

Hoberg and Phillips（2018）根据 Hoberg and Phillips（2016）的文本网络行业分类方法，重新检验了行业收益率的动量效应。他们发现相比于使用传统的行业分类，根据文本网络行业分类产生的行业收益率动量效应更加稳健并且具有强度大、持续时间长等特点。这是由于与按照传统 SIC 分类的行业内公司关联度相比，按照文本网络分类的行业内的公司关联程度可见性更低。因此，后者容易产生更严重的市场反应不足现象，从而导致长期并且显著的动量效应。他们进一步验证了这种动量效应可以被关注理论来解释。

3. 度量和预测经济周期

如何追踪和实时预测经济周期是经济学中的一个重要问题。由于衡量经济活动的主要变量 GDP 增长率无法实时观测，传统做法是使用市场上存在的一些即时指示变量，如金融市场、劳动力市场的数据等来作为反映经济活动的一致性指标。但这些方法的问题是，一方面这些指示变量和 GDP 增长率之间的关系不稳定，另一方面使用高频金融数据只能反映经济层面的一部分信息，很难判断是何种信息因素在影响或反映经济变动状况。与传统数据相比，新闻数据覆盖领域广泛、信息可以被很多经济个体所获取、并且新闻内容可能与经济当前和未来状态密切相关。基于这一思想，Thorsrud（2018）从挪威日度频率的商业新闻数据，结合季度 GDP 增长率数据，构建了日度经济周期指数。他们首先使用 LDA 模型从新闻数据中提取出了 80 个主题（财政政策、税收、货币政策等）。然后他们根据各个主题的语调（正面或负面），在混频时变动态因子模型的框架下，估计出了日度频率的新闻即时经济周期指数。他们发现，相比于使用现有的经济变量和一些复杂的经济模型，该指数能够更准确的预测和划分经济周期，并且样本外预测也有很高的准确性。另外，不同时期文章所包含的主题不同，因此对该经济周期指数分解出影响其波动的新闻类别，能够进一步推测出驱动或反应经济波动的因素。

除了构建经济周期之外，文献也有研究致力于考察媒体情绪和经济周期之间的关系。例如，Shapiro et al.（2018）使用 1980-2015 年美国 16 家主流新闻媒体的经济和金融相关的新闻数据，结合词典法和机器学习方法（商业公司提供的软件包）构建了反应经济状况的月度频率的情绪指数，该情绪指数包括负面、忧虑、满意等多个度量维度。他们研究了这些新闻

情绪与当前经济状态的相关性，以及这些情绪的变动对当前和未来经济状态的影响程度。他们发现新闻情绪中包含了能够预测未来经济状况的信息：媒体情绪与联邦基金利率、非农就业率、行业产出、实际个人消费支出等重要的经济周期指示变量存在很强的同期关系；并且对“通货膨胀率”和“联邦基金利率”等经济变量有预测能力。他们还发现新闻冲击和总需求冲击作用相似，即当媒体情绪变差的时候，就业率、通货膨胀率、联邦基金利率均会显著下降。

4. 媒体政治倾向

文本大数据有助于度量一些对经济政治生活比较重要、但过去无法量化的指标，如媒体政治倾向（Media Slant）。Gentzkow and Shapiro（2010）使用美国新闻报纸数据，根据媒体新闻语言与国会共和党和民主党语言的相似性，构建了媒体政治倾向指数。他们首先利用2005年国会议员发言记录数据，提取和民主党和共和党国会议员意识形态高度相关的1000个短语。最后将这些短语和政党意识形态对应起来，根据这1000个短语在议员报告中出现的相对频率和议员的政治形态，回归找出最能预测党派的特征性短语和相应回归系数。再将2000-2005年英文新闻头条数据按频率和词组长度挑选出新闻常用词汇，通过统计新闻报纸中这些短语出现的频率并结合回归系数对报纸的政治倾向进行分类。该方法的分类样本内的估计结果和真实分类结果的相关系数为0.61；而将分类出的媒体政治倾向和用户对这些报纸的政治倾向评级数据进行比对，发现两者的相关系数高达0.4。作者认为分类效果较好，并进一步研究媒体政治倾向与读者需求之间的关系。

（二） 在金融学中的应用

文本大数据在金融学的应用主要从度量关注度、情绪、隐含波动率和意见分歧，以及这四方面指标与市场表现的关系方面展开[12]。

1. 关注度指数

金融理论指出，关注是一种稀缺资源（Kahneman，1973），信息需要先被投资者关注到，才能通过投资者交易行为传递到资产价格中（Ben-Raphael et al.，2017），因此关注是信息反应的前提。从度量个体角度，关注度可分为投资者关注度（散户投资者和机构投资者）、媒体关注度和分析师关注度。由于现有文献对分析师关注的研究较少，本文主要梳理投资者关注和媒体关注相关应用。

（1）投资者关注度

散户和机构投资者是金融市场的直接参与者，研究他们的关注行为有助于理解资产价格的变动。Barber and Odean（2008）认为，由于购买股票时散户从他们关注的股票列表中作选择，但卖出股票时只能从持仓中选择，散户投资者关注度增加会导致暂时的价格上升。机构投资者持有更多股票、信息加工能力更强，因此通常不存在有限关注度约束。要检验Barber and Odean（2008）理论，关键是如何度量两类投资者的关注。

传统的关注度度量方法选择市场变量等作为关注度的代理变量，如交易量（Barber and Odean，2008；Peng and Xiong，2007）、超额收益率（Barber and Odean，2008）、广告费用（Lou，2014）等。但Da et al.（2011）指出，与投资者关注无关的因素也可以引发这些变量的变动。近年的研究开始直接用文本大数据构建散户关注度指标。

用文本数据度量散户投资者关注的方法主要有两类，一是利用网络搜索引擎统计对上市公司的搜索次数，二是网络论坛上股民对特定股票的发帖数量。Da et al.（2011）最早提

[12] 文本分析在金融研究中还有一个维度是文本的可读性（复杂性），常见文本可读性度量指标是迷雾指数（Fog Index）。Loughran and MacDonald（2014）发现可读性的度量还不如文本长短本身更有预测能力，这表明可读性相关研究还处于比较初步的阶段，因此本文不作详细介绍。

出并使用搜索次数度量投资者关注，他们根据谷歌趋势提供的搜索指数，使用 Russell 3000 成分股的代码作为关键字，构建了特定股票的投资者关注度。他们发现与 Barber and Odean（2008）的关注理论一致，高散户关注度预测了短期更高的收益率，但长期存在收益率反转。Antweiler and Frank（2004）则使用雅虎财经网络论坛的帖子数量来近似关注度，发现关注度对收益率和市场波动率均有预测能力，但对收益率的预测并不具有经济上的显著性。Tsukioka et al.（2018）使用雅虎财经日本板块上 654 家公司的帖子数据度量投资者关注度，并发现投资者关注可以解释日本上市公司的 IPO 抑价现象。

国内采用文本数据度量散户投资者关注的研究与国外做法类似，也是或者采用搜索指数或者使用论坛发帖量。在使用搜索指数方面，宋辉杰等（2011）使用中国 A 股 825 家上市公司的名称作为关键词，从谷歌趋势上获取这些公司的每周搜索量数据，并参照 Da et al.（2011）用周度异常搜索量来度量投资者关注，他们发现投资者关注可解释中国市场的 IPO 异象。俞庆进和张兵（2012）则才用百度搜索构建创业板 196 家公司个体投资者关注度，并发现中国创业板市场也存在投资者有限关注现象。张谊浩等（2014）则使用百度搜索指数研究了投资者网络搜索行为与资产定价间的关系，在关注度和短长期收益率以及交易量的关系方面，得到和 Da et al.（2011）在美国市场发现基本一致的结论。在使用网络论坛发帖量来度量散户投资者关注度的文献中，Huang et al.（2016）的发现是中国市场上投资者的关注具有本地偏好特征；杨晓兰等（2016）发现本地关注度与交易量存在正相关；段江娇等（2017）则发现帖子数与当日及未来的股票收益率显著负相关，但与当日及未来的股票波动率显著正相关。

由于缺乏直接反应机构投资者关注的文本数据，直接使用文本数据度量机构投资者关注的研究较少，Ben-Rephael et al.（2017）是首篇这类文献。通过分析 Bloomberg 的用户特征，他们发现 Bloomberg 使用者主要是机构投资者，并采用 Bloomberg 终端记录的用户对股票新闻的搜索和阅读频率数据来度量美国市场机构投资者关注度。他们发现与散户投资者关注相比，机构投资者关注对重大消息和事件反映更迅速，并且机构投资者关注领先于散户投资者关注。

（2）媒体关注度

媒体关注度（Media Coverage）反映的是媒体对于特定上市公司、行业或市场的关注程度，通常通过统计特定新闻媒体所发布的与金融市场、上市公司相关的新闻数量来构建。作为金融市场的信息制造和传播者，媒体的关注一方面可以影响市场参与者的关注，另一方面也影响市场信息的传播效率和模式。媒体关注对市场影响的研究，主要从它对资产价格、对管理层行为和分析师行为影响等角度展开。

从对资产价格影响的角度看， Fang and Peress（2009）选取了 1993-2002 年纽约时报、今日美国、华尔街日报和华盛顿邮报上关于 NYSE 和 NASDAQ 上市公司的新闻报道数据，从横截面研究了媒体关注与资产收益率的关系，并发现媒体关注低的公司的股票未来收益比媒体报道程度高的公司更高；Zou et al.（2018）发现中国股票市场上媒体关注和公司未来股票收益率之间也存在类似关系。Hillert et al.（2014）使用 1989-2010 年 45 家美国报纸约 220 万新闻数据研究了媒体关注与股票市场动量效应的关系，他们的发现可总结为受关注更高的公司的收益率可预测性更强，因此他们认为媒体关注会导致更严重的投资者偏差。

媒体关注对管理层行为影响方面，Dyck et al.（2008）使用俄罗斯 1999-2002 年的公司治理的违规数据，研究了媒体报道与公司违规行为的关系。他们发现金融时报和华尔街日报等国际媒体对违规事件的关注越高，公司纠正违规行为的概率越高。周开国等（2016）在中国市场上研究了媒体监督与上市公司违规频率之间的关系，也发现媒体关注度的提高会延长公司的违规间隔、降低违规频率。

对分析师的行为方面，周开国等（2014）发现媒体关注度可影响分析师关注度，从而提高其盈余预测的准确度。谭松涛等（2015）则发现媒体关注度能够降低分析师的预测乐观度和预测偏差。

2. 文本情绪

因为情绪的变化可能会导致资产价格偏离正常水平（Delong et al. 1990），度量情绪（sentiment）是文本大数据在金融领域的一大应用。文献中情绪常有正面和负面、乐观和悲观、积极和消极、牛市和熊市、看涨和看跌等不同表述，也常用“语调（tone）”来表示“情绪”。根据情绪主题的不同，文本情绪研究对象主要包括媒体语调（媒体新闻）、管理层语调（上市公司年报的管理层讨论与分析、盈利电话会议和其他公开信息披露文件）、投资者情绪（网络论坛发帖）等。情绪有关的文献主要从媒体情绪、管理层情绪和投资者情绪三个方面展开。

（1）媒体情绪（语调）

媒体情绪度量媒体报道内容中包含的乐观与悲观情绪。国外文献使用华尔街日报、纽约时报、华盛顿邮报等文本数据来度量媒体情绪，并研究媒体情绪与股票市场的关系。Tetlock（2007）研究了华尔街日报专栏文章和随后股票市场收益及交易量之间的关系，发现如果消极词语出现频率上升，则股票市场收益率会下降。Tetlock et al.（2008）则使用 1980-2004 年约 35 万篇华尔街日报和道琼斯新闻社上与标普 500 公司相关的新闻数据，发现公司新闻中负面词语比例越高，下个交易日公司股票收益率和下个季度公司盈利都更低。这表明新闻中的负面词汇包含了公司基本面难以量化的新信息。Garcia（2013）进一步拓展了 Tetlock（2007）的研究。他们使用 1905-2005 年纽约时报上的金融新闻，研究了经济繁荣期和衰退其媒体情绪对资产价格影响的不对称性。他们发现新闻在日度频率上对收益率的预测能力主要存在于经济衰退期。由于 Garcia（2013）还发现正面还是负面语调均能预测收益率，因此实际研究中应同时考虑文本中的正负语调相关词汇。

国内文献也用国内主流财经媒体报刊数据来度量媒体情绪，并考察媒体情绪与资产价格的关系。游家兴和吴静（2012）选取了 2004-2010 年国内 8 家主流财经报纸上的新闻，通过人工阅读新闻报道态度倾向的方法来衡量媒体情绪；他们发现媒体情绪越极端，沪深 A 股上市公司资产的错误定价越严重。汪昌云和武佳薇（2015）使用六家主流财经媒体的新闻数据，结合自定义的财经媒体情绪词典统计了新闻中的正负面词语数量并构建媒体正、负面语气指数。他们发现媒体负面语气能够解释 IPO 抑价率的变化，但正面语气却没有解释能力。

随着我国互联网金融行业的快速发展，国内文献还研究了媒体情绪与网络借贷之间的关系。王靖一和黄益平（2018）使用和讯网 1702 万余条新闻数据，度量了 2013 年 1 月至 2017 年 9 月间的金融科技情绪指数，用于反映媒体新闻对金融科技的正负情感态度。他们发现媒体情绪对于个体网络借贷具有显著影响，媒体情绪转向乐观时会增大网络借贷平台交易量的增长率，并且这种影响在问题平台上更强。张皓星和黄益平（2018）使用该指数进一步研究了互联网金融情绪与互金平台贷款违约率的关系。他们发现当金融科技情绪变差时，网络借贷违约概率会显著增加。

（2）管理层语调

除了使用财务报表直接报告公司经营状况外，上市公司还要定期发布公司的财务报告（季报、年报）季度盈余公告、管理层盈余公告、招股说明书等文件。这些文件包含上市公司管理层对当前经营状况的分析和未来发展方向的讨论，因此往往能反映管理层的决策和意图。

从这类文本数据中提取文本情绪并研究其对上市公司市场表现就属于对管理层语调的研究。例如，Li（2010）从美国上市公司年报（10-K）和季报（10-Q）文件的管理层讨论与分析（Management Discussion and Analysis，MD&A）前瞻性说明部分构建管理层语调，发

现这些语调与公司未来盈利存在正相关关系。Loughran and McDonald（2011）利用 1994-2008 年美国上市公司的年报文件，发现年报语调与收益率、交易量、波动率、未预期盈利等市场变量相关。Jegadeesh and Wu（2013）研究了 1995-2010 年上市公司的年报语调与年报发布期的超额收益率之间的关系，发现市场对于年报的内容存在反应不足的现象。他们还采用同样的方法度量了 IPO 招股说明书的语调，发现 IPO 招股说明书的语调与 IPO 抑价显著负相关。

国内文献从公司业绩、投资者交易行为等角度研究了管理层语调的影响。谢德仁和林乐（2015）使用 2005-2012 年中国上市公司年度业绩说明会的文本数据，发现业绩说明会中的管理层语调与未来公司的业绩存在显著正相关关系。曾庆生等（2018）研究了 2007-2014 年 A 股非金融公司年报语调与公司高管的交易行为，发现积极的年报语调预示公司高管随后的卖出股票规模大、净买入股票规模小。

（3）投资者情绪

实证检验投资者情绪与资产价格之间的关系首先需要测度投资者情绪。Delong et al.（1990）将投资者情绪定义为噪声交易者（Noise trader）关于股票未来股价偏离理性套利者信念的程度。传统的投资者情绪度量方法分市场变量法和调查法。Baker and Wurlger（2006）[13]的投资者情绪指数是目前文献中应用最广泛的基于市场变量法的投资者情绪指数。他们选取封闭式基金折价率、NYSE 股票换手率、IPO 数量及上市首日收益率、新发行权益份额和股利溢价作为情绪的代理变量来构建情绪指数。调查法则通过问卷调查（电话、邮件等）来收集个体对当前或未来经济状况、金融市场走向的看法和态度，并将这些问卷结果汇总成指数。密歇根大学的消费者信心指数[14]是调查法的经典代表。

这两种度量投资者情绪的方法各有弱点。市场变量法的问题是作为投资者情绪代理变量的市场变量可能不只反映投资者情绪，还反映情绪与其他经济因素相互作用后的均衡结果（Qiu and Welch，2006；Da et al.，2014）。调查法虽然旨在直接度量投资者情绪，但其实施成本高、构建情绪指数频率较低、时间跨度也比较短。

文本大数据为度量投资者情绪提供了新的数据源。一方面，由于投资者倾向于选择在网络论坛上发布与股票相关的评论帖子或者作出相关搜索，这些文本数据能直接反映他们对公司未来的看法、对市场当前状态的解读以及与自身投资决策相关的信息。另一方面，这些数据具有易获得、时间跨度长、覆盖公司数量多等特点，满足了从不同频率、不同层面研究情绪与资产价格关系的需求。

国外度量投资者情绪文献较为丰富，表一总结了其中的代表文献。其中度量投资者情绪的文本数据主要来源于雅虎财经帖子（Antweiler and Frank，2004；Kim and Kim, 2014；Tsukioka et al.，2018）、微博平台（Renault, 2017）、Twitter（Behrendt and Schmidt，2018）、专业数据库（Sun et al. 2016）和谷歌搜索（Da et al., 2014，Gao et al. 2018）。相应的，数据类型主要是帖子、推文或者搜索的关键词等。数据频率则涵盖了周频、日频、甚至更高频率（根据一分钟频率数据计算半小时频率的情绪指数）。利用构造出的文本情绪指数，文献研究了文本情绪和同期收益率、未来收益率、波动率、交易量、IPO 折价之间的关系。这些研究发现，在日度频率上，投资者情绪与同期收益率存在正相关，但基本对未来收益率、波动率、交易量没有很强的预测能力（（Antweiler and Frank，2004）。在更高的半小时频率上，一些研究发现了投资者情绪对收益率存在一定日内预测能力（如 Sun et al. 2016；Renault, 2017）；对于波动率的日内预测能力也有一定证据，但其经济意义不显著（Behrendt and Schmidt，2018）。目前为止，国外文献基本指向基于文本的投资者情绪的预测能力主要体现在对日内收益率的可预测性，而这一预测能力主要由于噪声交易者的交易行为导致的

[13] http://people.stern.nyu.edu/jwurgler/。
[14] http://www.sca.isr.umich.edu/。

（Sun et al. 2016）。美国股票市场比较成熟，并且以机构投资者为主、散户投资者占有比例远小于中国散户投资者占比。因此，散户投资者情绪对于市场表现的影响主要在日内体现。

表 1 投资者情绪相关研究

文献	数据来源	数据类型	情绪指数类型	主要发现
Antweiler and Frank（2004）	2000 年雅虎财经和 Raging Bull	道琼斯工业平均指数和道琼斯互联网指数的 45 家公司约 150 万帖子数据	个股层面投资者情绪指数	情绪与同期收益率显著正相关，与未来收益率不相关
Kim and Kim（2014）	2005-2010 年雅虎财经	91 家公司约 3200 万帖子	个股层面投资者情绪指数	情绪不能预测收益率、波动率、交易量，但受收益率影响
Tsukioka et al.（2018）	2001-2010 年雅虎财经日本股票版块	654 家公司相关的帖子数据		投资者情绪可用于解释 IPO 抑价现象
Sun et al.（2016）	1998-2011 年汤普森路透数据库	标普 500 指数对应的 1 分钟频率情绪数据		日内半小时情绪变化可预测日内收益率
Renault（2017）	2012-2016 年微博平台 StockTwits	约 6 千万帖子		第一半小时投资者情绪变化能预测标普 500 指数 ETF 最后半小时收益率，但下个交易日反转
Behrendt and Schmidt（2018）	2015-2017 年 Twitter	道琼斯指数成分股情绪数据（1 分钟频率）	个股 Twitter 情绪	情绪与日内波动率存在反馈效应，但经济意义不显著
Da et al.（2011）	谷歌搜索	118 词日搜索频率	FEARS 指数近似市场情绪	FEARS 能预测股票市场波动，收益率存在同期正相关，与随后收益率负相关
Gao et al.(2018)	2004-2014 年谷歌搜索	词搜索频率	周度国别投资者情绪指数	投资者情绪指数负向预测未来一周股票市场的收益率；并检验定价困难和有限套利等可预测性渠道

国内文献关于从文本数据度量投资者情绪的研究与国外研究类似，也是从网络平台数据出发，根据股民发布的帖子来构建中国股票市场的投资者情绪指数，并检验投资者情绪与市场变量的关系。杨晓兰等（2016）使用东方财富股吧上与创业板股票相关的约一年间 90

多万帖子构建了投资者情绪指数，并研究情绪指数与收益率、交易量等市场变量间的同期关系。段江娇等（2017）使用东方财富网 2011-2012 年上证 A 股约 466 万帖子构建了日度频率的投资者情绪指数，他们也发现论坛情绪与公司股票收益率存在同期相关，但并不能预测未来股票收益率。

简而言之，现有国内外文献都观察到投资者情绪与市场变量间的同期关系或者投资者情绪受市场变量影响，但投资者情绪的预测能力有限。这一现象一方面可能由于在有效的市场中，情绪对于市场变化的作用有限决定，另一方面也可能是由于情绪与市场变量同时变化的内生性导致低估了情绪的作用。Chen et al.（2018）则考虑利用中国股市开盘收盘时间段特征来建立因果关系。具体而言，他们使用 2008-2018 年中国股吧论坛的数据，结合词典法和机器学习方法构建了投资者隔夜情绪，即每个交易日收盘后到第二个交易日上午 9:15 分集合竞价前这一时段的投资者情绪。由于这段时间内只有投资者在论坛发表言论而没有价格信息，因此隔夜情绪与第二个交易日的开盘价之间只存在单向关系。他们发现隔夜投资者情绪能够显著预测隔夜收益率、第二天的日内收益率以及交易量。 根据他们的研究，网络论坛帖子反映的投资者情绪不只是噪音，还包含了关于未来市场的重要信息。

3. 新闻隐含波动率指数

除了用来度量媒体关注度和媒体语调外，新闻文本还被用来度量金融市场的不确定性。Manela and Moreira（2017）使用华尔街日报 1890-2009 年头版数据，采用支持向量回归法将新闻文本数据中出现的词语和市场上的波动率指数（VIX）相对应并构建了新闻隐含波动率指数（News implied volatility，NVIX）。

该指数的具体构建方法如下：（1）从每个月的所有新闻文章中提取全部词语出现的频率，采用独热法构建向量X_t，即X_t的长度为所有新闻中单词的个数，每个位置的值表示该词语在该月文章中出现的频率。（2）将X_t与 VIX 指数v_t构建映射：

$$v_t = w_0 + w \cdot X_t + v_t$$

（3）将 VIX 数据样本拆分为训练集和测试集，在训练集上使用支持向量回归方法拟合上述方程，得到系数的估计值；（4）根据每个月构建的新闻向量X_t，向前估计 NVIX 指数。

基于新闻数据构建的 NVIX 指数跨度从 1890 到 2009 年，该指数与历史上的重要事件（一战、二战、大萧条等）非常吻合，间接验证了该指数很好的刻画了市场的不确定性。他们发现 NVIX 可以正向预测市场 6 到 24 个月的收益率。进一步将新闻中的词语分为四类：股票市场、战争、政府、金融中介，他们发现 NVIX 预测收益率的能力主要受新闻中与战争、政府相关的词语的比例的影响。

NVIX 指数的构建思想还可以应用到其他文本数据上，通过选择不同的文本特征X_t，结合不同的市场变量v_t，包括超额收益率、交易量、波动率等，寻找这些文本特征跟市场变量之间的对应关系，提取更丰富的文本隐含信息。

4. 投资者分歧

投资者分歧衡量了投资者的异质信念，传统金融理论指出，投资者分歧会产生交易（Harris and Raviv，1993），因此文献关心分歧与交易量、价格之间的关系。常用的度量投资者分歧的指标包括分析师预测分散程度（Yu，2011）、经济不确定性指数（Bollerslev et al.，2018）和对经济变量预测的分散程度等。近年来，文献中开始从文本数据出发，构建直接度量投资者分歧的指标。

Antweiler and Frank（2004）使用网络留言板的帖子数据，计算出帖子的情绪得分，然后根据帖子情绪的标准差构建了投资者分歧指数。他们发现投资者分歧与同期的交易量显著正相关，验证了投资者分歧产生交易的理论。段江娇等（2017）在中国市场也有类似的发现。他们使用东方财富股吧帖子数据构建了日度频率的投资者分歧，发现投资者分歧越高，未来两天的交易量也越大。

投资者分歧对价格也会产生影响。Miller（1977）指出当市场上的投资者观点不同时，乐观的交易者会推动价格上升，而悲观的交易者由于存在卖空约束的限制，并不能完全消除由乐观交易者导致的错误定价，导致资产价格被高估。因此当投资者的分歧很大时，资产价格会被高估、未来收益率会更低。Hillert et al.（2018）使用 1989-2010 年美国主流媒体的新闻数据，先用词典法计算出每篇新闻的语调，再计算公司层面的媒体分歧程度（公司i在第t天的分歧程度为当天与该公司相关的全部新闻语调的标准差），最后将公司层面的媒体分歧程度加总并平均得到日度市场层面的媒体分歧程度指数。他们发现，媒体分歧程度与第二天的市场收益率显著负相关，并且这种关系在经济处于衰退时期更强。在公司层面，他们还发现媒体分歧对于 Beta、分析师预测分散程度、换手率、特质性波动率更高的公司的影响更大。

在本部分我们文本大数据在经济和金融领域的运用作了简单的梳理，自然不能穷尽目前日渐涌献的各个子领域的新文献，而一些文献也不能简单分类到一个子领域。例如，Soo（2018）使用 2000-2013 年美国 34 个城市的地方房地产行业新闻数据，采用词典法构建了房地产情绪指数，他们发现媒体新闻所反应的房地产行业情绪能够预测未来的房价变化。这是媒体情绪相关文献在经济学领域的运用。另外，手机、摄像机等产生的数据也应文本形式被用于研究中，如 Athey et al. (2018)采用手机产生的数据研究了消费者对餐厅的选择。当然，文本大数据在经济和金融外的其他领域，如社会学、政治学等领域也有精彩运用，可参考 Gentzkow et al. (2018)去参阅相关经典文献。

四、 结论和展望

和经典的数据分析方法相比，文本大数据给经济学和金融学的实证研究至少带来了四个变化。第一，经典实证分析采用的数据往往已经是结构化数据，而文本大数据往往是非结构化数据。非结构化数据向结构化数据转换的实现方式并不是一个简单的问题，而不同转换方式会直接影响后续分析结果，因此可以预见的是，高质量的文本大数据分析，需要对这一转换过程做更为详尽的介绍。第二，经典实证分析数据中的变量定义往往比较清晰，这种清晰的边界往往是通过收集数据时设计的问卷、或者实际经济和金融活动运行需要就已经事先界定来实现。而文本大数据的数据来源是新闻媒体、网络论坛、公司财报等文本文件，本身并不包含清晰的变量，如何提取信息、并且如何论证作者提取的就是目标信息，也将是文本大数据分析的重要步骤。第三，经典实证方法往往采用获取的全样本数据做回归分析，然后通过不同的设定和变化做稳健性检验。将数据分为训练集、检验集合测试集，展开交叉验证的方法虽然早就是经典方法，在过去的经济和金融实证分析中的应用还不够充分。由于这些方法在文本大数据做预测的分析范式中较为常见，这些做法对于经济和金融实证研究范式也会产生一定影响。第四，使用文本大数据需要有跨学科领域的人才。比如卷积神经网络、循环神经网络等模型在经济和金融领域的运用，需要研究人员不仅对于经济和金融领域有较为深入的掌握，同时对于不同算法的特点和优劣都要有较为丰富的知识。

未来几年，在经济和金融领域运用文本大数据研究方面，可能会有如下趋势。一是研究将开拓更为丰富的数据源。目前文本数据库的主要数据源包括新闻、网络论坛帖子、公司财报、消费者评价、重要人物的发言等。但还有大量的文本数据尚未被研究所使用，如政府工作报告和规划、网络自媒体公众号文章、微博大 V 观点、书籍、档案、专利网站、法院判决、医生处方等。运用有监督或者无监督的机器学习方法、深度学习方法来分析这些数据在未来几年将是研究热点。二是采用文本研究的问题会更为深入和广泛，如如何在中文语境以及复杂的长文本下的构建情绪指数并用于预测；如何从文本数据的文件大小、数字和汉字的

占比、图表数量、句子长短等角度加强对文本可读性的研究，又如如何提取国内主流新闻数据提取隐含信息（波动率、不确定性等），并考察文本信息与资本市场、宏观经济层面、微观个体表现等之间的关系。三是采用文本大数据展开的研究将不仅满足于基于相关关系的预测问题，因果关系相关的研究也将逐渐进入研究人员的视野。例如，Athey (2017)就考虑了如何将机器学习用到基于文本、摄像头等产生的数据而展开的政策效果的评估中。最后，由于文本大数据分析往往需要同时运用经济、金融、计算机、心理学等多个领域的知识和技术，对高素质的跨学科人才将产生较大需求，因此研究机构和高校可按照自身的学科优势来培养跨学科、复合型的研究人才。

参考文献

[1] Ackert, Lucy F., et al. "Influential investors in online stock forums." *International Review of Financial Analysis* 45 (2016): 39-46.

[2] Antweiler, Werner, and Murray Z. Frank. 2004 "Is all that talk just noise? The information content of internet stock message boards." *Journal of finance* 59.3: 1259-1294.

[3] Airoldi, E. M., Erosheva, E. A., Fienberg, S. E., Joutard, C., Love, T., and Shringarpure, S. (2010). Reconceptualizing the classification of PNAS articles. *Proceedings of the National Academy of Sciences*.

[4] Athey, Susan, David Blei, Robert Donnelly, Francisco Ruiz and Tobias Schmidt, 2018, “Estimating Heterogenous Consumer Preferences for Restaurants and Travel Time Using Mobile Location data”, arXiv:1801.07826.

[5] Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, *131*(4), 1593-1636.

[6] Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, *61*(4), 1645-1680.

[7] Barber, B. M., and Odean, T, “All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *Review of Financial Studies*, *21*(2008).

[8] Bali, Turan G., and Hao Zhou. "Risk, uncertainty, and expected returns." *Journal of Financial and Quantitative Analysis* 51.3 (2016): 707-735.

[9] Behrendt, Simon, and Alexander Schmidt. "The Twitter Myth Revisited: Intraday Investor Sentiment, Twitter Activity and Individual-Level Stock Return Volatility." *Journal of Banking & Finance* (2018).

[10] Ben-Rephael, A., Da, Z., & Israelsen, R. D. (2017). It depends on where you search: institutional investor attention and underreaction to news. *The Review of Financial Studies*, *30*(9), 3009-3047.

[11] Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford university press.

[12] Blei, David M., and John D. Lafferty. "Dynamic topic models." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.

[13] Brogaard, Jonathan, and Andrew Detzel. "The asset-pricing implications of government economic policy uncertainty." *Management Science* 61.1 (2015): 3-18.

[14] Bollerslev, Tim, Jia Li, and Yuan Xue. "Volume, volatility and public news announcements." *The Review of Economic Studies* (2018).

[15] Box, Travis. (2018), “Qualitative similarity and stock price comovement”, Journal of Banking

and Finance, 91, 49-69.
[16] Brown, Stephen V., and Jennifer Wu Tucker. (2011), "Large-sample evidence on firms' year-over-year MD&A modifications." *Journal of Accounting Research* 49.2, 309-346.
[17] 陈胜蓝、李占婷，“经济政策不确定性与分析师盈余预测修正”，《世界经济》，2017 年第 7 期，第 169-192 页。
[18] 陈胜蓝、刘晓玲，“经济政策不确定性与商业公司信用供给”，《金融研究》，2018年第5期，第172-190页。
[19] 陈霄、叶德珠、邓洁，“借款描述的可读性能够提高网络借款的成功率吗”，《中国工业经济》，2018 年第 3 期，第 174-192 页。
[20] Chen, Y., Huang Z., Li, J., Shen, Y. and Wang J. (2018) Can text-based investor sentiment help understand Chinese stock market? A deep learning method. Working Paper.
[21] Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *The Journal of Finance*, *66*(5), 1461-1499.
[22] Da, Z., Engelberg, J., & Gao, P. (2014). The sum of all FEARS investor sentiment and asset prices. *The Review of Financial Studies*, *28*(1), 1-32.
[23] Das, Sanjiv R., and Mike Y. Chen. "Yahoo! for Amazon: Sentiment extraction from small talk on the web." *Management science* 53.9 (2007): 1375-1388.
[24] Davis, Angela K., Jeremy M. Piger, and Lisa M. Sedor. "Beyond the numbers: Measuring the information content of earnings press release language." *Contemporary Accounting Research* 29.3 (2012): 845-868.
[25] De Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann, "Noise trader risk in financial markets." *Journal of political Economy* 98.4 (1990): 703-738.
[26] Dyck, Alexander, Natalya Volchkova, and Luigi Zingales. "The corporate governance role of the media: Evidence from Russia." *The Journal of Finance* 63.3 (2008): 1093-1135.
[27] 段江娇、刘红忠、曾剑平，“中国股票网络论坛的信息含量分析”，《金融研究》，2017 年第 10 期，第 178-192 页。
[28] Elman, Jeffrey L. "Finding structure in time." *Cognitive science*14.2 (1990): 179-211.
[29] Fang, Lily, and Joel Peress. "Media coverage and the cross-section of stock returns." *The Journal of Finance* 64.5 (2009): 2023-2052.
[30] Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, *15*(4), 915-953.
[31] Gao, Zhenyu, Haohan Ren, and Bohui Zhang. "Googling investor sentiment around the world." *Journal of Financial and Quantitative Analysis* (2018) forthcoming.
[32] Garcia, Diego. "Sentiment during recessions." *The Journal of Finance* 68.3 (2013): 1267-1300.
[33] Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. *Text as data*. No. w23276. National Bureau of Economic Research, 2017.
[34] Gentzkow, Matthew, and Jesse M. Shapiro, ‘‘What Drives Media Slant? Evidence from US Daily Newspapers,’’ Econometrica, 78 (2010), 35–71
[35] Gulen, Huseyin, and Mihai Ion. "Policy uncertainty and corporate investment." *The Review of Financial Studies* 29.3 (2015): 523-564.
[36] Gunning, Robert, 1952, The Technique of Clear Writing (McGraw-Hill, New York).
[37] 顾研、周强龙，“政策不确定性、财务柔性价值与资本结构动态调整”，《世界经济》，2018 年第 6 期，第 102-126 页。
[38] Harris, Milton, and Artur Raviv. "Differences of opinion make a horse race." *The Review of*

Financial Studies 6.3 (1993): 473-506.
[39] Henry, Elaine. "Are investors influenced by how earnings press releases are written?." *The Journal of Business Communication (1973)* 45.4 (2008): 363-407.
[40] Hillert, Alexander, Heiko Jacobs, and Sebastian Müller. "Media makes momentum." *The Review of Financial Studies*27.12 (2014): 3467-3501.
[41] Hillert, Alexander, Heiko Jacobs, and Sebastian Müller. "Journalist disagreement." *Journal of Financial Markets*(2018).
[42] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *science*313.5786 (2006): 504-507.
[43] Hirshleifer, David, and Siew Hong Teoh. "Limited attention, information disclosure, and financial reporting." *Journal of accounting and economics* 36.1-3 (2003): 337-386.
[44] Hoberg, G. and G. M. Phillips (2016). Text-based network industries and endogenous product differentiation. Journal of Political Economy 124(5), 1423–1465.
[45] Hoberg, Gerard, and Gordon Phillips. "Text-based industry momentum." *Journal of Financial and Quantitative Analysis* (2018)， forthcoming.
[46] Huang, Allen H., Amy Y. Zang, and Rong Zheng. "Evidence on the information content of text in analyst reports." *The Accounting Review* 89.6 (2014): 2151-2180.
[47] Huang, Yuqin, Huiyan Qiu, and Zhiguo Wu. "Local bias in investor attention: Evidence from China's Internet stock message boards." *Journal of Empirical Finance* 38 (2016): 338-354.
[48] Huang, Z., C. Tong, H. Qiu, and Y. Shen, "The spillover of macroeconomic uncertainty between the US and China." *Economics Letters*171 (2018): 123-127.
[49] Jegadeesh, Narasimhan, and Di Wu. "Word power: A new approach for content analysis." *Journal of Financial Economics* 110.3 (2013): 712-729.
[50] Jurado, Kyle, Sydney C. Ludvigson, and Serena Ng. "Measuring uncertainty." *American Economic Review* 105.3 (2015): 1177-1216.
[51] 纪洋、王旭、谭语嫣、黄益平，“经济政策不确定性、政府隐性担保与企业杠杆率分化”，《经济学（季刊）》，2018 年第 2 期，第 449-470 页。
[52] Kahneman, Daniel. *Attention and effort*. Vol. 1063. Englewood Cliffs, NJ: Prentice-Hall, 1973.
[53] Kearney, Colm, and Sha Liu. "Textual sentiment in finance: A survey of methods and models." *International Review of Financial Analysis* 33 (2014): 171-185.
[54] Kim, Soon-Ho, and Dongcheol Kim. "Investor sentiment from internet message postings and the predictability of stock returns." *Journal of Economic Behavior & Organization* 107 (2014): 708-729.
[55] Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).
[56] Lai, S., L. Xu, K. Liu, and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification." *AAAI*. Vol. 333. 2015.
[57] Lawrence, Alastair. "Individual investors and financial disclosure." *Journal of Accounting and Economics* 56.1 (2013): 130-147.
[58] LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep learning. Nature 521(7553), 436–444.
[59] Li, Feng. "Annual report readability, current earnings, and earnings persistence." *Journal of Accounting and economics*45.2-3 (2008): 221-247.
[60] Li, Feng. "The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach." *Journal of Accounting Research* 48.5 (2010):

1049-1102.
[61] 李凤羽、杨墨竹，“经济政策不确定性会抑制企业投资吗？—基于中国经济政策不确定指数的实证研究”，《金融研究》，2015 年第 4 期，第 115-129 页。
[62] Lou, Dong. "Attracting investor attention through advertising." *The Review of Financial Studies* 27.6 (2014): 1797-1829.
[63] Loughran, Tim, and Bill McDonald. "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks." *The Journal of Finance* 66.1 (2011): 35-65.
[64] Loughran, Tim, and Bill McDonald. "Measuring readability in financial disclosures." *The Journal of Finance* 69.4 (2014): 1643-1671.
[65] Loughran, Tim, and Bill McDonald. "Textual analysis in accounting and finance: A survey." *Journal of Accounting Research* 54.4 (2016): 1187-1230.
[66] Manela, Asaf, and Alan Moreira. "News implied volatility and disaster concerns." *Journal of Financial Economics* 123.1 (2017): 137-162.
[67] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).
[68] Miller, Brian P. "The effects of reporting complexity on small and large investor trading." *The Accounting Review* 85.6 (2010): 2107-2143.
[69] Miller, Edward M. "Risk, uncertainty, and divergence of opinion." *The Journal of finance* 32.4 (1977): 1151-1168.
[70] Murphy, K. P. (2012). Machine Learning: a Probabilistic Perspective. MIT Press
[71] 孟庆斌、杨俊华、鲁冰，“管理层讨论与分析披露的信息含量与股价崩盘风险-基于文本向量化方法的研究”，《中国工业经济》，2017 年第 12 期，第 132-150 页。
[72] Naik, Nikhil, Ramesh Raskar and César A. Hidalgo. 2016. "Cities Are Physical Too: Using Computer Vision to Measure the Quality and Impact of Urban Appearance." *American Economic Review*, 106(5): 128-32.
[73] Peng, Lin, and Wei Xiong. "Investor attention, overconfidence and category learning." *Journal of Financial Economics* 80.3 (2006): 563-602.
[74] 彭俞超、韩珣、李建军，“经济政策不确定性与企业金融化”，《中国工业经济》，2018 年第 1 期，第 137-155 页。
[75] Doran, James S., David R. Peterson, and S. McKay Price. "Earnings conference call content and stock price: the case of REITs." *The Journal of Real Estate Finance and Economics*45.2 (2012): 402-4
[76] Qiu, Lily, and Ivo Welch. *Investor sentiment measures*. No. w10794. National Bureau of Economic Research, 2004.
[77] Renault, Thomas. "Intraday online investor sentiment and return patterns in the US stock market." *Journal of Banking & Finance* 84 (2017): 25-40.
[78] Rogers, Jonathan L., Andrew Van Buskirk, and Sarah LC Zechman. "Disclosure tone and shareholder litigation." *The Accounting Review* 86.6 (2011): 2155-2183.
[79] Shapiro, Adam Hale, Moritz Sudhof, Daniel Wilson, 2018. “Measuring News Sentiment”, Federal Reserve Bank of San Francisco Working Paper 2017-01.
[80] Solomon, David H., Eugene Soltes, and Denis Sosyura. "Winners in the spotlight: Media coverage of fund holdings as a driver of flows." *Journal of Financial Economics* 113.1 (2014): 53-72.

[81] Soo, Cindy K. "Quantifying Sentiment with News Media across Local Housing Markets." *The Review of Financial Studies* (2018) forthcoming.
[82] 宋辉杰、曹晖、杨坤，“投资者关注与 IPO 异象——来自网络搜索量的经验证据”，《经济研究》，2011 年增 1 期，第 145-155 页。
[83] Sun, Licheng, Mohammad Najand, and Jiancheng Shen. "Stock return predictability and investor sentiment: A high-frequency perspective." *Journal of Banking & Finance* 73 (2016): 147-164.
[84] Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* (2006): 101(476), 1566–1581.
[85] Tetlock, Paul C. "Giving content to investor sentiment: The role of media in the stock market." *The Journal of finance* 62.3 (2007): 1139-1168.
[86] Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy. "More than words: Quantifying language to measure firms' fundamentals." *The Journal of Finance* 63.3 (2008): 1437-1467.
[87] Thorsrud, Leif Anders. "Words are the New Numbers: A Newsy Coincident Index of the Business Cycle." *Journal of Business & Economic Statistics,* (2018): 1-35.
[88] Tsukioka, Yasutomo, Junya Yanagi, and Teruko Takada. "Investor sentiment extracted from internet stock message boards and IPO puzzles." *International Review of Economics and Finance* 56 (2018): 205-217.
[89] Wang,Jingyi, Yan Shen, Yiping Huang, “How Does News Sentiment Affect the Peer-to-Peer Lending Market in China?”, (2018): mimeo.
[90] 谭松涛、甘顺利、阚铄，“媒体报道能够降低分析师预测偏差吗”，《金融研究》，2015 年第 5 期，第 192-206 页。
[91] Vapnik, V. (1996). The Nature of Statistical Learning Theory. New York: Springer.
[92] 汪昌云、武佳薇，“媒体语气、投资者情绪与 IPO 定价”，《金融研究》，2015 年第 9 期，第 174-189 页。
[93] 王红建、李青原、刑斐，“经济政策不确定性、现金持有水平及其市场价值”，《金融研究》，2014 年第 9 期，第 53-68 页。
[94] 王靖一、黄益平，“金融科技媒体情绪的刻画与对网贷市场的影响”，《经济学（季刊）》，2018 年第 4 期，第 1623-1650 页
[95] 谢德仁、林乐，“管理层语调能预示公司未来业绩吗？-基于我国上市公司年度业绩说明会的文本分析”，《会计研究》，2015 年第 2 期，第 20-27 页
[96] 杨晓兰、沈翰彬、祝宇，“本地偏好、投资者情绪与股票收益率：来自网络论坛的经验证据”，《金融研究》，2016年第12期，第143-158页。
[97] 游家兴、吴静，“沉默的螺旋：媒体情绪与资产误定价”，《经济研究》，2012年第7期，第141-152页。
[98] Yu, Jialin. "Disagreement and return predictability of stock portfolios." *Journal of Financial Economics* 99.1 (2011): 162-183.
[99] 俞庆进、张兵，“投资者有限关注与股票收益——以百度指数作为关注度的一项实证研究”，《金融研究》，2012 年第 8 期，第 152-165 页
[100] 曾庆生、周波、张程、陈信元，“年报语调与内部人交易：表里如一还是口是心非？”，《管理世界》，2018 年。
[101] 张皓星、黄益平，“情绪、违约率与方向挤兑——来自某互金企业的证据”，《经济学（季刊）》，2018 年第 4 期，第 1503-1524 页

[102] 张谊浩、李元、苏中锋、张泽林，“网络搜索能预测股票市场吗？”，《金融研究》，2014 年第 2 期，第 193-206 页

[103] 周开国、应千伟、陈晓娴，“媒体关注度、分析师关注度与盈余预测准确度”，《金融研究》，2014 年第 2 期，第 139-152 页。

[104] 周开国、应千伟、钟畅，“媒体监督能够起到外部治理的作用吗？-来自中国上市公司违规的证据”，《金融研究》，2016年第6期，第193-206页。

[105] Zou, Liping, Kien Dinh Cao, and Yishun Wang. "Media Coverage and the Cross-Section of Stock Returns: The Chinese Evidence." *International Review of Finance* (2017)